Nowcasting influenza epidemics using non homogenous hidden Markov models

Baltazar Nunes[*], Departmento de Epidemiologia,
Instituto Nacional de Saúde Dr. Ricardo Jorge, Portugal

Isabel Natário, CEAUL, Departmento de Matemática,
Faculdade de Ciências e Tecnologia,
Universidade Nova de Lisboa, 2829-516, Caparica, Portugal

M. Lucília Carvalho, CEAUL, Departamento de Estatística e
Investigação Operacional,
Faculdade de Ciências, Universidade de Lisboa, Portugal

August 17, 2011

Abstract

Timeliness of a public health surveillance system is one of its most important characteristics. The process of predicting the present situation using available incomplete information from surveillance systems

[*]Address: Departamento de Epidemiologia, Instituto Nacional de Saúde Dr.Ricardo Jorge, Av. Padre Cruz, 1649-016 Lisboa, Portugal. Telephone: +351217526490 Email: baltazar.nunes@insa.min-saude.pt

has received the term of *nowcasting* and has high public health interest. Generally in Europe the epidemiological surveillance of the influenza activity is supported by general practitioners sentinel networks, and the epidemiological bulletins of each week are usually issued between Wednesday and Friday of the following week.

In this work we develop a non homogenous hidden Markov model (HMM) that, in a weekly basis, uses as covariates an early estimate of influenza-like illness (ILI) incidence rate and the number of ILI cases tested positive to nowcast the current week ILI rate and the probability that the influenza activity is in an epidemic state.

Bayesian inference is used to obtain estimates of the model parameters and nowcasted quantities. The results obtained by application to data provided by the Portuguese influenza surveillance system, show the additional value of using a non homogenous HMM instead of an homogenous one. It is possible to demonstrate that using a non homogenous HMM improves the surveillance system timeliness in 2 weeks.

Keywords: hidden Markov models, nowcasting, surveillance, influenza

1 Introduction

During public health threats like infectious diseases outbreaks (SARS, West Nile virus, pandemic influenza, etc) knowing in real-time the epidemics trends, spatial distribution and impact in terms of medical consultations, hospitalizations and deaths is essential to identify the most appropriate

measures to control the disease spread and mitigate its impact[1, 2, 3]. It is widely accepted that the most proper tool to acquire reliable information on disease spread and impact remains in the use of stable disease surveillance systems defined as *"...the ongoing, systematic collection, analysis, interpretation and dissemination of disease related data for public health action..."*[1].

Generally in Europe the epidemiological surveillance of the influenza activity is supported by sentinel systems, formed by general practitioners, that provide weekly the number of consultations with patients presenting influenza-like illness (ILI) symptoms. These figures divided by the number of patients in the GP list or the total number of consultations on that week enable the estimation of ILI incidence rates. On other hand GP sentinel networks also provide, for some of the ILI patient reported, nasopharyngeal swabs that after laboratory confirmation support the virological surveillance of the influenza activity[4, 5].

At the European level the influenza surveillance system is managed by the European Center for Disease Prevention and Control (ECDC) through a network of countries named European Influenza Surveillance Network. In this framework each Wednesday all participating countries upload age-specific ILI rates and virological information corresponding to the previous week[1] in a web-based system (TESSy). The European influenza activity report

[1]ISO definition: Monday through Sunday

gathering information from all the participating countries (WISO: weekly influenza surveillance overview) is issued by the ECDC on the following Friday[5].

In a no delay reporting situation one could consider that by Monday all the information to estimate ILI rates for the previous week should be complete. So for a particular week t, ILI rate estimated on Monday of week $t+1$ could be considered the zero day delay ILI rate estimate. Having made these considerations, in the actual European surveillance system, countries upload age-specific ILI rates with a 2 days delay and ECDC issues the WISO with a 4 days delay.

Timeliness of a public health surveillance system is one of its most important characteristics, given that it is crucial for its capacity of a timely intervention[1, 2]. For this study, timeliness will be considered as the time elapsed from the disease onset to the generation of an automated alert.

The prospective detection of the beginning of the epidemic period has been done by a variety of statistical methods such as regression techniques, time series methods, methods of statistical process control and also on statistical multivariate analysis using multiple data sources. A comprehensive review of these methods can be found in [6, 7, 8, 9]. Some of these methods were implemented in the R package `surveillance`[10] and can be easily used for ongoing disease surveillance. Nevertheless, for all these methods the main goal is to identify the first week of the epidemic period, when the ILI in-

cidence rate indicates levels of influenza activity that can be classified as epidemic.

Regarding the influenza surveillance systems based on GP sentinel networks, all these methods were only applied to data referring to the previous week. This means that the online detection algorithms could only report an alert at most by the beginning of the following week. Despite that, data providers (GP) send the data to the surveillance system on a daily basis. This has became more promptly since some surveillance systems use web-based systems where GP can enter data during consultation with the patient [11] or use computer routines to capture data from GP electronic medical records[12, 13]. Therefore this daily data stream could capacitate the surveillance systems with sufficient information to assess the current situation, without time delay, enabling the real-time early detection of the epidemic start, peak and end.

The process of predicting the present situation using the available incomplete information has been considered of high interest by public health officials, mainly during the pandemic (H1N1)2009 receiving the term of *nowcasting* [14]. The use of incomplete information was already applied to nowcast on a weekly basis, the number of influenza A(H1N1)2009 hospitalizations during the 2009-10 pandemic in the Netherlands with considerable sucess[15].

In this work we developed a statistic model that in a weekly basis, uses all data collected by a surveillance system before the end of the week, to

nowcast two measures of interest: the ILI incidence rate and the influenza activity state, epidemic or non-epidemic, that will be reported by Wednesday of the following week to ECDC. Given this objective, we were able to show the adding value of using a non homogenous hidden Markov model (HMM), being the advantage mainly shown in its ability of using covariates, with early information on the epidemic (like weekly ILI cases laboratory confirmed, early estimates of the ILI rate, ILI rate for other age groups, etc), to model the influenza activity state transition probabilities. In opposition to the homogenous model used in prior studies [16, 17, 18], where the state transition probabilities remains the same. Additionally we were also capable of introducing covariates in the response variable (Wednesday ILI rate) in order to nowcast its specific value. To our knowledge this work represents the first attempt to use non homogenous HMMs in a disease surveillance problem with the objective of early detect an outbreak and nowcast its evolution.

The present article is organized in the following way. In section 2 the application of the HMMs to influenza surveillance is introduced and reviewed, the non homogenous model and the generic application to the subject of influenza surveillance and nowcasting are presented. Section 3 introduces the data used for the application example, i.e. the Portuguese influenza surveillance system from week 40/2008 to week 16/2011. In Section 4 the specific models proposed are described along with the bayesian estimation

approach for the model parameters and to nowcast the Wedenesday ILI rate and influenza activity state. Section 5 details the results, both the application of the models to the entire data set from week 40/2008 to week 16/2011 and the real-time nowcast of the 2010-11 influenza season. Finally in section 6 the model and results are discussed and the main conclusions presented.

2 Hidden Markov models in influenza surveillance

Generally a HMM assumes that the observed time series, y_t with $t = 1, ..., T$, is a realization of a stochastic process $\{Y_t : t = 1, ..., T\}$, where the distribution of each Y_t is conditionally determined by an unobserved discrete stochastic process $\{S_t : t = 1, ..., T\}$, that assumes values in a m-states set $\{1, 2, ..., m\}$. For the homogenous case, this unobserved stochastic process is assumed to be an order one Markov chain with stationary transition probabilities given by $\gamma_{j,i} = P(S_t = i|S_{t-1} = j)$ for any $i, j \in \{1, ..., m\}$.

The HMM were first applied to ILI surveillance data in 1999 by Le Strat and Carrat [16]. In their work they proposed a two states homogenous HMM, epidemic and non-epidemic, where in each of these states the weekly ILI rate was described by a normally distributed cyclical model of period 52 plus a linear trend. Following this first work others have emerged using the same line of approach, from which we point out the work of Rath et al 2003 [17], that proposed an i.i.d exponential distribution for rates in the non-epidemic state and an i.i.d normal distribution for the rates in the epidemic state,

and the work of Martinez-Beneito et al 2008 [18], that modeled the one week lag differences of the weekly ILI rates, considering that those for the non-epidemic state were normally distributed with mean zero, and those for the epidemic state were described by an order one autoregressive model.

As stated before the application of these models to ILI data had always the goal to classify weeks as epidemic or non epidemic using all available data[16, 17] or, like in Martinez-Beneito et al 2008 [18], to do this in an online basis, classifying the last week available as epidemic or non-epidemic. This last paper has used a bayesian approach for the estimation.

It is important to notice that these approaches did not have the objective of predicting or forecasting the ILI rates or even to forecast the hidden Markov chain state for the forthcoming weeks.

Given our objective of nowcasting the current week ILI rate (not yet observed) and the respective influenza activity state using the incomplete data collected by the influenza surveillance system, we need a model that enables the use of covariates with forecasting capacity. For this purpose we propose the use of a HMM that allows the introduction of covariates to model the weekly ILI rate and the sate transition probabilities. This late innovation implies the use of a non homogenous HMM instead of a homogenous one as in the previous literature[16, 17, 18], where the transition probabilities from the non-epidemic to the epidemic state (the epidemic beginning) and from the epidemic to the non-epidemic state (the end of the epidemic) were

invariant in time.

More specifically, we propose the application of a HMM were the state transition probabilities are given by a time-dependent matrix Γ^t with elements $\gamma^t_{j,i} = P(S_t = i|S_{t-1} = j)$ for any $i, j \in \{0, 1\}$ and $t = 2, ..., T$, where 0 and 1 represent, respectively, the non-epidemic and the epidemic state of influenza activity. Formally:

$$\Gamma^t = \begin{bmatrix} \gamma^t_{0,0} & \gamma^t_{0,1} \\ \gamma^t_{1,0} & \gamma^t_{1,1} \end{bmatrix},$$

Choosing the transition probabilities to be modeled by a logistic function of the covariates[19, 20],

$$logit(\gamma^t_{j,i}) = \ln\left(\frac{\gamma^t_{j,i}}{\gamma^t_{j,j}}\right) = \boldsymbol{\alpha}_{j,i}\boldsymbol{Z}_t$$

$$\gamma^t_{j,i} = \frac{\exp(\boldsymbol{\alpha}_{j,i}\boldsymbol{Z}_t)}{(1 + \exp(\boldsymbol{\alpha}_{j,i}\boldsymbol{Z}_t))}$$

for any $i, j \in \{0, 1\}$ and $j \neq i$, where $\boldsymbol{Z}_t = (1, z_{1,t}, ..., z_{q,t})$ is the vector of the q covariates measured in time t. The non homogeneity of the state transition probabilities can preclude the stationarity of the Markov chain[19], this can happen mainly when the covariates used are functions of time. In this situation an initial distribution for the non homogenous hidden Markov chain is defined by $\boldsymbol{\delta} = (\delta_0, \delta_1)$ that will be expressed has $\delta_i = P(S_1 = i)$ with $i \in \{0, 1\}$.

To better understand the advantages of the non homogenous HMM we apply this model to the data collected by the Portuguese Influenza Surveillance

system. For comparison purposes we also apply to the same data an equivalent model with a homogenous hidden Markov chain.

3 Data description

In Portugal the clinical component of the influenza surveillance system is assured by a network of 144 voluntary general practitioners ("Medicos Sentinela") that, since 1989, weekly report the ILI cases[2] that occurred among their patient list[21, 11] to the surveillance system hub (since year 2000, the Department of Epidemiology of National Health Institute Dr. Ricardo Jorge), by regular mail or using a web-based questionnaire. The population under observation, constituted by the total of the GP patients lists, represents a sample fraction of 2.5% of the total population.

The laboratorial component is coordinated by the National Influenza Reference Laboratory and consists in receiving swabs from ILI cases sent by a fraction of GPs from the sentinel system and from the network of emergency departments from hospitals and health centers for virological surveillance purposes.

The Portuguese system, as part of the EISN, calculates and reports each Wednesday to TESSy system, the age-specific ILI incidence rates and the number of ILI cases with laboratory confirmed influenza infection corre-

[2]ILI definition used by the ICPC - WONCA

sponding to the week before[3]. This is done because there are delays in the system, namely the time between the patient onset of symptoms to the GP visit and the GP reporting time to the surveillance system.

In order to have a predictor of the ILI rate reported to ECDC by Wednesday, the Portuguese influenza surveillance system has been calculating the age-specific ILI incidence rates for each week with the incomplete data gathered on Friday of that same week, from season 2008-2009 to 2010-2011. Our main goal is to access if there is enough information in this variable used as predictor for an early detection of the epidemic start, peak and end.

In synthesis, the data used in this study is the weekly ILI incidence rate (per 100,000 inhabitants) of week t calculated by Friday of week t, referred as $y_{t(t)}$, and the weekly ILI rate calculated by Wednesday of week $t+1$, referred as $y_{t(t+1)}$ (Figure1), for the period from week 40 of 2008 to week 16 of 2011. Virological data consists of the weekly number of ILI cases with laboratory confirmation for the influenza virus corresponding to week t, but calculated at Wednesday of week $t+1$: $v_{t(t+1)}$.

4 Models

The main objective is to nowcast the influenza activity state and the Wednesday ILI rates for week t, calculated in week $t+1$, using available data from

[3]The date of reference is the disease onset date

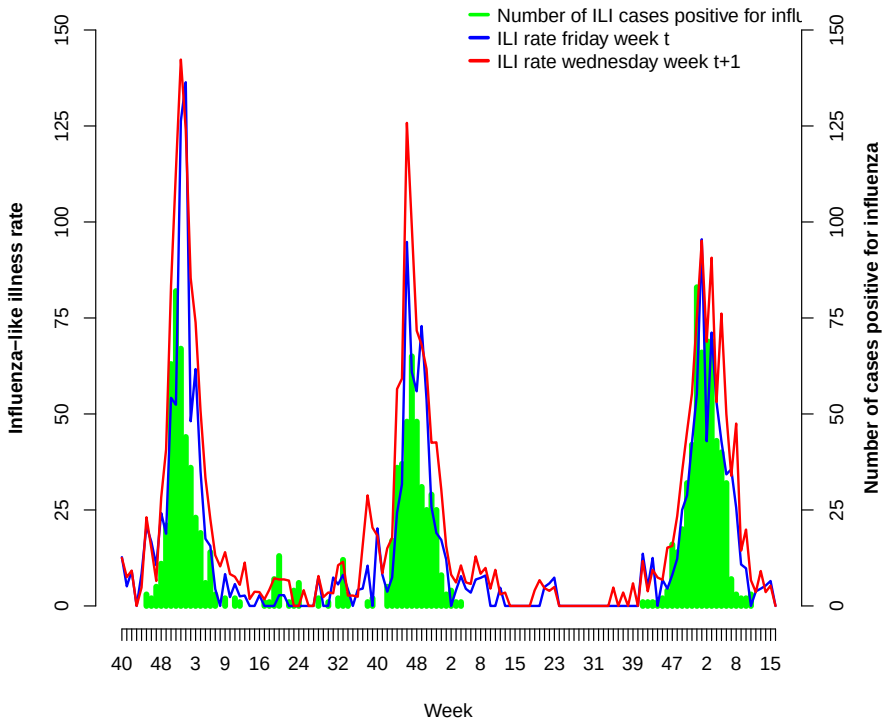

Figure 1: Influenza-like illness incidence rates calculated by Friday of week t and by Wednesday of week $t+1$.

covariates at time t, namely the early estimate of the ILI rates by Friday and the number of ILI cases tested positive for influenza for the previous week, estimated by Wednesday of the present week, $v_{t-1(t)}$. To achieve this goal a HMM, based on the work developed by Paroli and Spezia 2008 [20], that allows the inclusion of covariates in response variable and in the transition state probabilities is proposed.

From our observation, the influenza epidemic is sustained when the number

of ILI cases tested positive for influenza in a certain week is high, e.g. 20. On the other hand if the number ILI case confirmed for influenza is zero, the influenza activity state is clearly non epidemic. Based on this, if an increase in the ILI rate is observed without ILI cases confirmed for influenza, one can not assume that this increase is due to an influenza epidemic, but can be related to the circulation of other respiratory viruses. Using this empirical thresholds, two covariates that are functions of the early estimate of the ILI rate and of the number of ILI cases positive for influenza in the previous week are proposed. The objective of these two covariates is to enhance the capacity of the models to discriminate an observed ILI rate as belonging to either the epidemic or non epidemic period, keeping its prediction ability: one covariate to model the response variable in the non epidemic period,

$$y_{t(t)0} = \begin{cases} y_{t(t)} & v_{t-1(t)} \leq \nu_0 \\ 0 & \text{otherwise} \end{cases}, \tag{1}$$

and the other to model the response variable in the epidemic period,

$$y_{t(t)1} = \begin{cases} y_{t(t)} & v_{t-1(t)} \geq \nu_1 \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

In this study ν_0 was set at 20 and ν_1 to 1. As can be seen both variables share a common part of the early estimate of the ILI rate $y_{t(t)}$ i.e. the weeks where the number of ILI cases tested positive, in the previous week, was

higher or equal to 1 and lower or equal to 20.

Baring this in mind, the weekly incidence rate calculated by Wednesday $y_{t(t+1)}$ is assumed to be described by the following equations, conditional on the influenza activity state:

$$y_{t(t+1)} = \begin{cases} \mu + \beta_1 \cos(\frac{2\pi t}{52}) + \beta_2 \sin(\frac{2\pi t}{52}) + \theta_{0,1} y_{t(t)0} + e_{t,0} & S_t = 0 \\ \mu + \beta_1 \cos(\frac{2\pi t}{52}) + \beta_2 \sin(\frac{2\pi t}{52}) + \theta_{1,1} y_{t(t)1} + \theta_{1,2} y_{t(t)1}^2 + e_{t,1} & S_t = 1 \end{cases} \quad (3)$$

were $e_{t,i} \sim N(0, \tau_i)$, with precision $\tau_0 > \tau_1$, $t = 1, ..., T$ and $i \in \{0, 1\}$. Given that the epidemic state is characterized by values of ILI rate higher than the ones in the non epidemic state, a constraint that obliges the variance of the epidemic state to be higher than the variance of the non epidemic state is included in the model.

In this model we propose a common component for the epidemic and non-epidemic period that reflects the baseline behavior of the ILI rates. This component is constituted by a common mean μ and a cyclical component of period 52 weeks. The difference between the equations for each state is set on the association with the predictor, i.e. the state specific variables that are functions of the early estimate of the ILI rate on Friday of week t, presented in equation 1 and 2. So for the non-epidemic model a linear association with $y_{t(t)0}$ is considered, and on other hand the epidemic period is described by a quadratic association with $y_{t(t)1}$. We propose these two state dependent equations, from the observation of Figure 2 and with the rational that during the non epidemic period, i.e. for weeks with zero or a small number of

ILI cases tested positive, the number of ILI cases are uniformly distributed within the week, so the relation between the early estimate of ILI rate by Friday and the ILI rate estimated by Wednesday of the following week is linear. On the other hand during the epidemic period, the distribution of the ILI cases within each week will not be uniform, presenting a sharp growth or decrease, respectively, in the beginning and at the end of the epidemic. Given this, we propose that for the epidemic period the association between the two ILI rates estimates is quadratic.

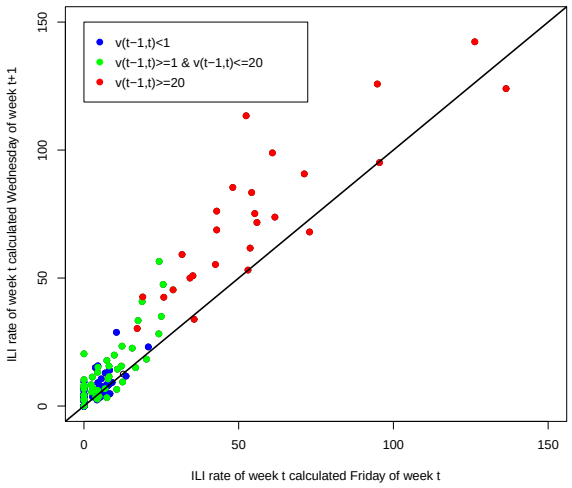

Figure 2: Association between ILI incidence rates calculated by Friday of week t and by Wednesday of week $t+1$ according to the number of ILI cases tested positive for influenza in the previous week $v_{t-1(t)}$. Black line represents $y_{t(t+1)} = y_{t(t)}$

For the hidden Markov chain state transition probabilities a time-varying

matrix with elements $\gamma_{j,i}^t$ for any $i, j \in \{0, 1\}$ and $t = 2, ..., T$ was considered. For a specific week t, $\gamma_{0,1}^t$ and $\gamma_{1,1}^t$ represent the probability that in week t the influenza activity is epidemic given that in the week before the influenza activity was respectively in the non-epidemic state or in the epidemic state. Given these considerations, three models are proposed (Models 0, 1 and 2). All three share the same model for the response variable, expressed in equation 3, but have different choices for the state transition probabilities. Models 1 and 2 have a non-homogenous hidden Markov chain, that are differentiated according to the covariates used in the logistic function that models the transition probabilities:

- Model 1:

$$logit(\gamma_{j,i}^t) = \ln\left(\frac{\gamma_{j,i}^t}{\gamma_{j,j}^t}\right) = \alpha_{j,i,0} + \alpha_{j,i,1} y_{t(t)}$$

$$\gamma_{j,i}^t = \frac{\exp(\alpha_{j,i,0} + \alpha_{j,i,1} y_{t(t)})}{(1 + \exp(\alpha_{j,i,0} + \alpha_{j,i,1} y_{t(t)}))}$$

- Model 2:

$$logit(\gamma_{j,i}^t) = \ln\left(\frac{\gamma_{j,i}^t}{\gamma_{j,j}^t}\right) = \alpha_{j,i,0} + \alpha_{j,i,1} y_{t(t)} + \alpha_{j,i,2} v_{t-1(t)}$$

$$\gamma_{j,i}^t = \frac{\exp(\alpha_{j,i,0} + \alpha_{j,i,1} y_{t(t)} + \alpha_{j,i,2} v_{t-1(t)})}{(1 + \exp(\alpha_{j,i,0} + \alpha_{j,i,1} y_{t(t)} + \alpha_{j,i,2} v_{t-1(t)}))}$$

for any $j, i \in \{0, 1\}$ and $j \neq i$.

More specifically, for Model 1, the transition probabilities from the non-epidemic state to the epidemic state and *vice versa* are modeled by the early estimate of the ILI incidence rate. In Model 2 transition probabilities are

modeled by the early estimate of ILI incidence rate and by the absolute number of ILI cases tested positive for influenza in the week before.

Finally for Model 0 a homogenous hidden Markov chain is used, where $\gamma_{i,j}^t = \gamma_{i,j}$ for any $t = 2, ..., T$ and for any $j, i \in \{0, 1\}$, with the objective of evaluating the additional value of a time-varying hidden Markov chain.

4.1 Parameters and hidden states estimation

The model parameters and the hidden states $\boldsymbol{\Psi} = (\mu, \boldsymbol{\tau}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{s}^T)'$ for the non-homogenous models 1 and 2, and $\boldsymbol{\Psi} = (\mu, \boldsymbol{\tau}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{s}^T)'$ for the homogenous model 0 are numerically estimated using a bayesian approach via Markov chain Monte Carlo (MCMC) methods. For the homogenous model 0, all the parameters are sampled using the Gibbs algorithm, for the non-homogenous models 1 and 2 exception is made for the parameters $\boldsymbol{\alpha}$ of the time-variant transition probabilities that are sampled using a Metropolis-Hastings algorithm. The state sequence $\boldsymbol{s}^T$ is sampled using the *ff-bs* algorithm: forward filtering - backward sampling algorithm[19, 20, 22] (see Appendix 1 for a detailed description of the algorithm used).

The initial distribution of the hidden Markov chain is fixed as an uniform discrete distribution ($\delta_0 = \delta_1 = \frac{1}{2}$). The parameter independent prior distributions are set as:

- for all the models:
  - $\mu \sim N(\mu_M, \sigma_M^2)$ where $\mu_M = 50$ and $\sigma_M^2 = 10$ given that, em-

pirically the rate of 50 ILI cases per 100.000 was the value above which Public Health officials in Portugal considered the start of the epidemic when baseline approaches were absent;

- $\tau_i \sim Gamma(\alpha_\Sigma; \beta_\Sigma)$, where $\alpha_\Sigma = \beta_\Sigma = 0.5$, under the increasing order constraint ($\tau_0 > \tau_1$), for $i = 0, 1$;
- $\theta_0 \sim N(\mu_\theta; \sigma^2_\theta)$ where $\mu_\theta = 0$ and $\sigma^2_\theta = 10$;
- $\boldsymbol{\theta}_1 \sim N_2(\boldsymbol{\mu}_\theta; \Sigma_\theta)$ where $\boldsymbol{\mu}_\theta = (0, 0)$ and $\Sigma_\theta = 10I_2$;
- $\boldsymbol{\beta} = (\beta_1, \beta_2) \sim N_2(\boldsymbol{\mu}_B; \Sigma_B)$ where $\boldsymbol{\mu}_B = (0, 0)$ and $\Sigma_B = 5I_2$;

- for Model 0: $\boldsymbol{\gamma}_0 = (\gamma_{0,0}, \gamma_{0,1})$ and $\boldsymbol{\gamma}_1 = (\gamma_{1,0}, \gamma_{1,1}) \sim Diriclet(\lambda_1, \lambda_2)$ where $\lambda_1 = \lambda_2 = 1$;
- for Model 1: $\boldsymbol{\alpha}_{0,1} = (\alpha_{0,1,0}, \alpha_{0,1,1})$ and $\boldsymbol{\alpha}_{1,0} = (\alpha_{1,0,0}, \alpha_{1,0,1}) \sim N_2(\mu_A; \Sigma_A)$ for $j, i \in \{0, 1\}$, where $\mu_A = (0, 0)$ and $\Sigma_A = 10I_2$;
- for Model 2: $\boldsymbol{\alpha}_{0,1} = (\alpha_{0,1,0}, \alpha_{0,1,1}, \alpha_{0,1,2})$ and $\boldsymbol{\alpha}_{1,0} = (\alpha_{1,0,0}, \alpha_{1,0,1}, \alpha_{1,0,2}) \sim N_3(\mu_A; \Sigma_A)$, where $\mu_A = (0, 0, 0)$ and $\Sigma_A = 10I_3$;

For the non homogenous models the posterior distribution of $\boldsymbol{\Psi}$ is then given by:

$$\pi\left(\boldsymbol{\Psi}|\boldsymbol{y}^T_{(+1)}, \boldsymbol{y}^T_0, \boldsymbol{C}, \boldsymbol{Z}, \boldsymbol{\delta}\right) = \pi\left(\mu, \boldsymbol{\tau}, \theta_0, \boldsymbol{\theta}_1, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{s}^T|\boldsymbol{y}^T_{(+1)}, \boldsymbol{y}^T_{(0)0}, \boldsymbol{y}^T_{(0)1}, \boldsymbol{C}, \boldsymbol{Z}, \boldsymbol{\delta}\right) \quad (4)$$

$$\propto f\left(\boldsymbol{y}^T_{(+1)}|\mu, \boldsymbol{\tau}, \theta_0, \boldsymbol{\theta}_1, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{s}^T, \boldsymbol{y}^T_{(0)0}, \boldsymbol{y}^T_{(0)1}, \boldsymbol{C}\right) f\left(\boldsymbol{s}^T|\boldsymbol{\alpha}, \boldsymbol{Z}, \boldsymbol{\delta}\right) p(\mu)p(\boldsymbol{\tau})p(\theta_0)p(\boldsymbol{\theta}_1)p(\boldsymbol{\beta})p(\boldsymbol{\alpha})$$

where $\boldsymbol{y}^T_{(+1)} = (y_{1(2)}, ..., y_{T(T+1)})'$ is the vector of the incidence rates estimated by Wednesday of week $t + 1$, $\boldsymbol{s}^T = (s_1, ..., s_T)$ is the vector of

the hidden states of Markov chain, $\boldsymbol{y}_{(0)0}^{T} = (y_{1(1)0}, ..., y_{T(T)0})'$ and $\boldsymbol{y}_{(0)1}^{T} = (y_{1(1)1}, ..., y_{T(T)1})'$ are the state specific vectors of the early estimate of the incidence rate calculated by Friday of week t, $\boldsymbol{C} = (\boldsymbol{c}_1, \boldsymbol{c}_2)$ is $2 \times T$ matrix of the periodic component $\boldsymbol{c}_1 = (\cos(\frac{2\pi}{52}), ..., \cos(\frac{2t\pi}{52}), ..., \cos(\frac{2T\pi}{52}))'$ and $\boldsymbol{c}_2 = (\sin(\frac{2\pi}{52}), ..., \sin(\frac{2t\pi}{52}), ..., \sin(\frac{2T\pi}{52}))'$ and $\boldsymbol{Z} = \boldsymbol{y}_{(0)}^{T}$ for Model 1 and $\boldsymbol{Z} = (\boldsymbol{y}_{(0)}^{T}, \boldsymbol{v}_{(+1)}^{T-1})$ for Model 2 are respectively the vector and matrix of the covariates included in state transition probabilities matrix, where $\boldsymbol{v}_{(+1)}^{T-1} = (v_{0(1)}, ..., v_{T-1(T)})'$.

The likelihood can be factorized as:

$$f\left(\boldsymbol{y}_{(+1)}^{T} | \mu, \boldsymbol{\tau}, \theta_0, \boldsymbol{\theta}_1, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{s}^T, \boldsymbol{y}_{(0)0}^{T}, \boldsymbol{y}_{(0)1}^{T}, \boldsymbol{C}\right) =$$

$$= \prod_{t=1}^{T} f(y_{t(t+1)} | \mu, \boldsymbol{\tau}, \theta_0, \boldsymbol{\theta}_1, \boldsymbol{\beta}, s_t, y_{t(t)0}, y_{t(t)1}, c_{1,t}, c_{2,t})$$

Here $f(y_{t(t+1)} | \mu, \boldsymbol{\tau}, \theta_0, \boldsymbol{\theta}_1, \boldsymbol{\beta}, s_t, y_{t(t)0}, y_{t(t)1}, c_{1,t}, c_{2,t})$ is given by:

$$\sqrt{\frac{\tau_0}{2\pi}} \exp\left\{-\frac{\tau_0}{2}\left(y_{t(t+1)} - \mu - \beta_1 c_{1,t} - \beta_2 c_{2,t} - \theta_0 y_{t(t)0}\right)^2\right\} \qquad \text{when} \qquad s_t = 0 \qquad \text{and}$$

$$\sqrt{\frac{\tau_1}{2\pi}} \exp\left\{-\frac{\tau_1}{2}\left(y_{t(t+1)} - \mu - \beta_1 c_{1,t} - \beta_2 c_{2,t} - \theta_{1,1} y_{t(t)1} - \theta_{1,2} y_{t(t)1}^2\right)^2\right\} \qquad \text{if} \qquad s_t = 1.$$

Finally the joint distribution of the hidden states is given by:

$$f(\boldsymbol{s}^T | \boldsymbol{\alpha}, \boldsymbol{Z}, \boldsymbol{\delta}) = \delta_{s_1} \prod_{t=2}^{T} \gamma_{s_{t-1}, s_t}^{t}$$

For the homogenous HMM (Model 0) the state transition probabilities $\gamma_{i,j}^t = \gamma_{i,j}$, i.e. are obviously not indexed in time.

Finally the posterior probability that a week t belongs to the epidemic state is given by the posterior mean of each s_t, being estimated as:

$$\hat{P}[S_t = 1] = \sum_{k=1}^{K} \frac{s_t^{(k)}}{K}$$

where $s_t^{(k)} \in \{0, 1\}$ are the sampled states in each iteration and K is the total number of iterations of the MCMC algorithm, after burn-in period and thinning. The classification of each week in the epidemic on non epidemic state is named as state decoding.

4.2 Nowcasting weekly influenza activity states and ILI rates

In order to simulate the online performance of the models, the influenza activity state and the correspondent ILI rate of each week, from week 40/2010 to week 16/2011, were nowcasted. Each proposed model was sequentially fitted to all the data known until the respective previous week to nowcast the following. For example, week 40/2010 was nowcasted from week 40/2008 to week 39/2010, week 41/2010 was nowcasted from week 40/2008 to week 40/2010, and so on.

More specifically consider that we are on Friday of week T, knowing all ILI rates and ILI cases tested positive for influenza until week $T-1$, and we want to nowcast week T, using previous information and the incomplete estimate of the ILI rate of week T. The probability that week T belongs to

the epidemic influenza activity state is estimated by:

$$\hat{P}[S_T = 1] = \sum_{k=1}^{K} \frac{\hat{P}[S_T = 1|\Psi^{(k)}]}{K} \tag{5}$$

where $\hat{P}[S_T = 1|\Psi^{(k)}] = \hat{P}[S_{T-1} = 1|\Psi^{(k)}]\gamma_{1,1}^{T(k)} + \hat{P}[S_{T-1} = 0|\Psi^{(k)}]\gamma_{0,1}^{T(k)}$, k is the iteration step of the MCMC algorithm, $\hat{P}[S_{T-1} = 1|\Psi^{(k)}]$ is the probability that week $T-1$ belongs to the epidemic state sampled in the kth iteration by the *ff-bs* algorithm and $\gamma_{j,i}^{T(k)}$ with $j, i \in \{0, 1\}$ represent the transition probabilities sampled in the kth iteration of the MCMC algorithm. For model 0 the elements of transition probabilities sampled in each iteration are not time-dependent $\gamma_{j,i}^{(k)}$. On the other hand, for the non homogenous models the transition probabilities are estimated using the known covariates at moment T and the parameters $\boldsymbol{\alpha}_{j,i}^{(k)}$ sampled in iteration k:

$$\gamma_{j,i}^{T(k)} = \frac{\exp(\boldsymbol{\alpha}_{j,i}^{(k)} Z_T)}{(1 + \exp(\boldsymbol{\alpha}_{j,i}^{(k)} Z_T))}$$

were $Z_T = (1, y_{T(T)})$ for model 1 and $Z_T = (1, y_{T(T)}, v_{T-1(T)})$ for model 2. The nowcasting of the ILI rate for week T to be reported on week $T+1$ is estimated by:

$$\hat{y}_{T(T+1)} = \sum_{k=1}^{K} \frac{y_{T(T+1)|s_T^{(k)}=1}^{(k)} \hat{P}[S_T = 1|\Psi^{(k)}] + y_{T(T+1)|s_T^{(k)}=0}^{(k)} (1 - \hat{P}[S_T = 1|\Psi^{(k)}])}{K} \tag{6}$$

were $y_{T(T+1)|s_T^{(k)}=1}^{(k)}$ is sampled from $f(y_{T(T+1)}|\mu^{(k)}, \boldsymbol{\beta}^{(k)}, \tau_0^{(k)}, \theta_0^{(k)}, y_{T(T)0}, c_{1,T}, c_{2,T})$ and $y_{T(T+1)|s_T^{(k)}=0}^{(k)}$ is sampled from $f(y_{T(T+1)}|\mu^{(k)}, \boldsymbol{\beta}^{(k)}, \tau_1^{(k)}, \boldsymbol{\theta}_1^{(k)}, y_{T(T)1}, c_{1,T}, c_{2,T})$.

5 Results

5.1 Application to all data set

In a first phase all models were applied to the entire time series, from week 40/2008 to week 16/2011. This procedure had the objective of comparing models regarding the parameters estimates and the retrospective classification of each week in one of the Markov chain states (epidemic or non-epidemic), i.e. the decoding of the influenza activity states. Parameters and hidden states were estimated by a MCMC run of 200.000 iterations with a burn-in of 60.000 and a thinning of 100, for the non homogenous models, and 100.000 iterations with a burn-in of 25.000 and a thinning of 50, for the homogenous one. All results presented in this article were obtained using specific programs implemented in R computing language[23].

The MCMC output convergence was evaluated by the observation of the trace and auto-correlation functions of the parameters runs, and by the application of the statistic of Gelman-Rubin 1992 [24] and by the Raftery and Lewis method 1992 [25]. For this purpose the R package `coda` was used [26]. For all the parameters, the Gelman-Rubin scaling factors 97.5% percentiles were bellow 1.1, on the other hand the Raftery and Lewis method applied to the 1500 runs after the burn-in and thin suggested for all the parameters a burn-in not superior to 2 and a maximum number of iterations close 1500. These results indicate that convergency has been achieved enabling

the computations of the posterior means and credible intervals.

The Bayes factor was used in order to compare the models fit to data. For this purpose marginal likelihoods were computed using the method presented by Chib 1995 [27], for the homogenous model, given that all parameters are sampled using the Gibbs algorithm, and the method presented by Chib and Jeliazkov 2001 [28] for the non homogenous models, since the transition probabilities parameters are sampled via the Metropolis-Hastings algorithm. In Table 1 it can be seen that the non homogenous models present a better fit to data and, among these, model 2 does better.

According to Table 2 the estimates of parameters of the response variable

Model	**ln(mL)**
M0	-332.384
M1	-320.111
M2	-300.397

Table 1: Natural logarithm of the marginal likelihoods of the proposed models.

are very similar between the three models.
The only parameters that seem to change (decreasing in value) with the increase of model complexity, i.e. the introduction of covariates to model the state transition probability matrix, are the parameters of the common part that has the objective of explaining ILI rate baseline behavior (μ, β_1 and β_2).

Likewise the results obtained by Martinez-Beneito 2008 [18], in the homogenous model a week is more likely to belong to the epidemic state if the previous week was in the epidemic state ($\gamma_{1,1}$ posterior mean 0.91). This conservative result is also observed in the non-epidemic period, so a week is more likely to be in the non-epidemic state if the previous week was in the non-epidemic state ($\gamma_{0,0}$ posterior mean 0.95). This result also means that in each week the mean posterior probabilities of entering in the epidemic state and also leaving the epidemic state are constant over time and very low (respectively $\gamma_{0,1} = 0.05$ and $\gamma_{1,0} = 0.09$). For the non homogenous models (Figure 3), and as it was expected, the posterior probabilities of changing the influenza activity state in each week vary over time. Nevertheless some of the posterior 95% credible intervals for the transition probabilities matrix parameters include zero (see Table 2). Exception are observed for the probability of entering the epidemic state, in the constant ($\alpha_{0,1,0}$), that is significant in both models, and for the probability of leaving the epidemic state, in the parameters associated with the early estimate of ILI rate ($\alpha_{1,0,1}$), in model 1, and the one associated with number of ILI cases positive for influenza in the previous week($\alpha_{1,0,2}$), in model 2.

Figure 4 depicts, for all the models, the weekly mean posterior probability of being in the epidemic sate. This value is concentrated in the neighborhood of zero or one, meaning that there is a low proportion of weeks with doubtful classification, i.e. with the posterior mean probability of being in

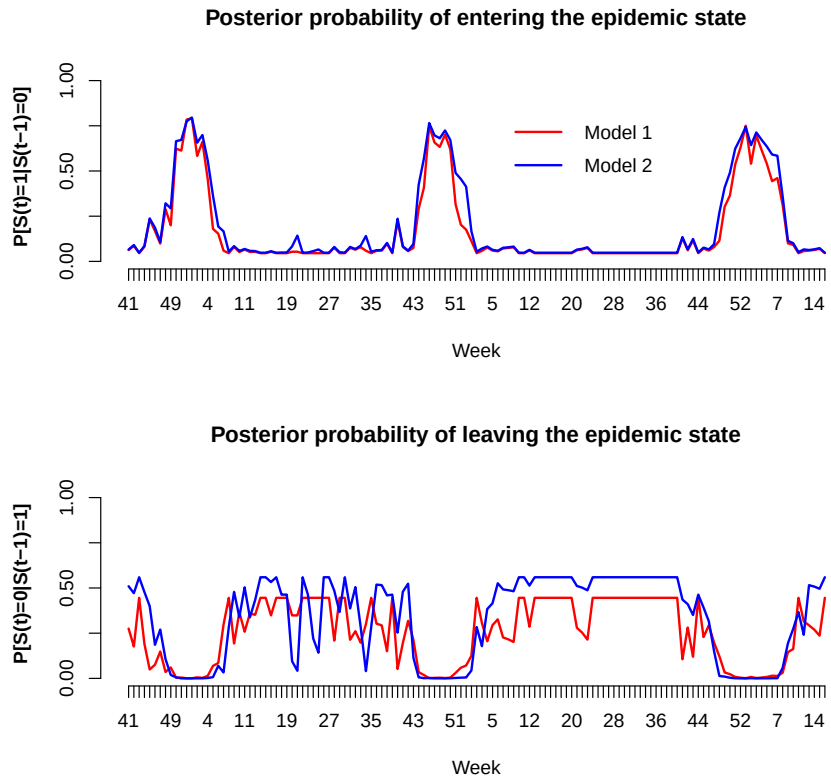

Figure 3: Mean posteriori probabilities of entering and leaving the epidemic influenza activity state according to the non-homogenous models (1 and 2).

the epidemic state near 0.5. Concerning the classification of each week in the epidemic or on the non-epidemic state, a week was considered as epidemic (non epidemic) if the posteriori probability of belonging to the epidemic state was higher or equal to 0.5 (lower than 0.5). An epidemic period was defined as the consecutive set of weeks with the mean posterior probability of being in the epidemic state above 0.5.

The non homogenous models identified four epidemic periods in the study

period while the homogenous model identified three (Table 3). In Figure 4 it can be seen that in season 2009-2010, that corresponds to the (H1N1)2009 pandemic in the period from week 37/09 to week 2/10, the non homogenous models identified two distinct epidemics but the homogenous one considered only one epidemic period that started by week 37/09 and ended by week 53/09. On the other hand, the epidemic periods for seasons 2008-09 and 2009-10 were consistently estimated by the three models.

5.2 Real-time nowcast of 2010-11 influenza season

In Figure 5 the results of the weekly real-time influenza activity state nowcast and decoding are presented for season 2010-11. In the first panel one can see the mean posterior probability (mpp) of being in the epidemic state calculated in the same week using equation 5. The second panel shows the weekly mpp of being in the epidemic state calculated in the following week, i.e., with the ILI rate of the interest week totally observed. Finally the last panel presents the weekly mpp of being in the epidemic state calculated after all the ILI rate values for season 2010-11 are know, i.e. by week 17/2011.
At the end of the season, by week 17/2011, all the three models define the epidemic period from week 46/2010 to week 10/2011, presenting very similar epidemic state mpp. In general, regarding the nowcasting and the decoding of the last ILI rate observed, the non-homogenous models presented an early

increase and decrease in the weekly epidemic state mpp, specially model 2.

If one considers 0.5 as the mpp cut-off value to classify a week in the epidemic period, the first signal of the epidemic start is observed in week 48/2010 when model 2 decodes ILI rate of week 47/2010 as epidemic and nowcast for week 48/2010 a epidemic state mpp very close to 0.5. This tendency continues to be observed in week 49/2010, where model 2 calculates an epidemic state mpp in the neighborhood of 0.5 for the observed week 48/2010 ILI rate and nowcast for that week (49/2010) a epidemic state mpp clearly higher than 0.5. These results are then confirmed in week 50/2011 by all the three models after week 49/2010 ILI rate became known. It is important to underline that the non-homogenous model 2 detects the signal of epidemic start two week before the homogenous model. On the other hand, looking for the signal of the end of the epidemic state, the non-homogenous models also show better results, given that epidemic state mpp start to decrease earlier. Nevertheless it is only by week 13/2011 that model 2 nowcasts the end of the epidemic, what is confirmed in week 14/2011 when week 13/2011 ILI rate is finally known. The homogenous model is clearly more slowly in entering and leaving the epidemic state, which tends to increase the timeliness in detecting the epidemic start and end.

On the other hand the weekly ILI rates nowcasted by the three models (obtained by equation 6) do not present relevant differences, although non-homogenous models present higher ILI rate estimates than the homogenous

model. As can be seen in Figure 6, weekly ILI rates of season 2010-11 were predicted during the same week in a very satisfactory way, they start to increase, reached the peak and decrease in synchrony. This means that the models were able to tackle the ILI rate evolution by Friday of the same week, reducing reporting delay in 4 days.

6 Discussion and Conclusion

Considering the main objective of this work, the models here proposed were able to nowcast the ILI incidence rate and the influenza activity state by Friday of the same week. These results were better achieved with the non homogenous HMM. As can be seen in the previous section, the non homogenous models were able to nowcast the beginning of the epidemic state two weeks before the homogenous model. On the other hand at the end of the epidemic, the probability of being in the epidemic state decreased more rapidly when calculated by the non homogenous models. Both these results underline the adding value of using non homogenous HMM to nowcast influenza epidemics. The inclusion of covariates, with early information on the epidemic evolution, to model the transition probabilities is of particulary importance to empower the model in the nowcast task. In true, using a homogenous HMM does not add too much to the nowcasting task given that the probability of a change is immutable in time.

Also important to notice is the fact that when the models were fitted to

the entire data set, the non homogenous models presented a better fit to the entire data set than the homogenous models, and identified in season 2009-10 (the pandemic (H1N1)2009) two distinct epidemic periods, where the homogenous model only identified one. During the 2009-10 influenza pandemic most of the European countries experienced two epidemic waves, one during Summer and a second one during Autumn. Portugal was not exception, in fact looking at other sources of information used during the pandemic (H1N1)2009 (e.g National Network of Laboratories for the A(H1N1) Diagnostic) a first small epidemic wave was identified in the period from week 31/2009 to week 39/2009 [29, 30]. All these features show that the non homogenous models give more reliable results.

Comparing both non homogenous models, model 1 and 2, model 2 presented the best results given that it produced the highest value of natural logarithm of the marginal likelihood and showed the higher capacity to nowcast the epidemic start in the real-time nowcast of season 2010-11.

Although these satisfactory results, the models proposed and data used have some caveats that must be addressed.

First the models were build using some ad-hoc decisions, more precisely regarding the constitution of the state specific covariates used to model the response variable $y_{t(t)0}$ and $y_{t(t)1}$ and also when establishing, respectively a linear and quadratic relation between this estimates and the reported ILI rate $y_{t(t+1)}$. In the first decision we set the cut-offs at $v_0 = 1$ and $v_1 = 20$

based on the empirical knowledge of the surveillance system. To evaluate the importance of this decision we have also fitted to all the data set the homogenous model using other cut-off values for v_1, more specifically 5 and 10, without substantial changes in the results. Nevertheless, to apply these models to other surveillance systems, these fixed parameters should be tuned using empirical knowledge or exploratory data analysis.

Regarding the decision to set the association of the early estimate of the ILI rate as linear for the non epidemic state and quadratic for the epidemic state, other options, namely linear-linear, were also tested in the homogenous models presenting worst results, mainly in the real-time nowcast of the ILI rate.

Other point of interest, but that works against the non homogenous models, is the higher number of iterations, burn-in and thin needed for the model MCMC output to converge. This fact has particular impact in the computational time needed to obtain the nowcast results each week. The main reason for this is that the covariates transition probabilities parameters $\boldsymbol{\alpha}$ are sampled using the random-walk Metropolis-Hastings algorithm, and this algorithm has a slower convergence than the Gibbs algorithm. In this study we have used a logistic function to model these probabilities, which unable the direct sampling of $\boldsymbol{\alpha}_{0,1}$ and $\boldsymbol{\alpha}_{0,1}$ from the posterior full conditional distribution. A possible further development for the proposed model could be to choose a function to model the transition probabilities that could enable

the use of the Gibbs algorithm to sample their specific parameters. Nevertheless this fact does not influence the timeliness of the nowcast objective, since the parameters estimation, that is more time consuming can be obtain previously (by Wednesday) and the nowcast by Friday when the early estimate of the ILI rate is available.

To conclude, the present work showed the advantage of using a non homogenous HMM to nowcast the influenza-like illness incidence rate and the influenza activity state in the context of an public health surveillance system. Additional we have also demonstrated that in our surveillance system the incomplete information from a GP based influenza surveillance enabled the early detection of the epidemic start. More specifically, it was possible to show that using a non homogenous HMM, with an early estimated of the ILI rate by Friday of the same week, improved the surveillance system timeliness in 2 weeks. Since the proposed models were fitted to data that corresponds to three influenza seasons of the Portuguese influenza surveillance system, and the real-time nowcast simulation was tested in one influenza season (2010-11), it is important to state that further applications of these models, to a higher number of seasons and also to data from other public health surveillance systems, are needed in order to warrant the adding value of using non homogenous HMMs to nowcast an epidemic evolution in the public health surveillance setting.

7 Acknowledgements

The authors would like to acknowledged the work of "Médicos-Sentinel" GP network for voluntarily providing weekly data for the influenza surveillance.The National Influenza Reference Laboratory of the National Health Institute Dr. Ricardo Jorge (INSA) for providing the data from the laboratorial component. All the colleagues of the Department of Epidemiology of the INSA, with a special thanks to Dr. José Marinho Falcão, Zilda Pimenta and Isabel Batista. This work is financed by National Funds through FCT - Fundação para a Ciência e a Tecnologia - in the scope of project PEst-OE/MAT/UI0006/2011.

	Model 0		Model 1		Model 2	
Parameter	Mean	95% CI	Mean	95% CI	Mean	95% CI
τ_0	0.134	[0.091;0.187]	0.137	[0.093;0.189]	0.139	[0.091;0.196]
τ_1	0.006	[0.004;0.009]	0.006	[0.004;0.009]	0.006	[0.004;0.009]
μ	3.131	[2.191;4.120]	3.011	[1.995;3.993]	2.876	[1.845;3.950]
β_1	-0.861	[-1.660;-0.048]	-0.799	[-1.633;0.004]	-0.720	[-1.517;0.133]
β_2	1.947	[0.753;3.216]	1.861	[0.617;3.122]	1.604	[0.289;2.954]
θ_0	0.710	[0.526;0.879]	0.726	[0.554;0.888]	0.739	[0.571;0.902]
$\theta_{1,0}$	1.556	[1.353;1.759]	1.571	[1.375;1.777]	1.576	[1.375;1.779]
$\theta_{1,1}$	-0.005	[-0.007;-0.003]	-0.005	[-0.007;-0.003]	-0.005	[-0.007;-0.003]
$\gamma_{0,0}$	0.948	[0.888;0.986]	NA	NA	NA	NA
$\gamma_{0,1}$	0.052	[0.014;0.112]	NA	NA	NA	NA
$\gamma_{1,0}$	0.092	[0.025; 0.191]	NA	NA	NA	NA
$\gamma_{1,1}$	0.908	[0.809;0.975]	NA	NA	NA	NA
$\alpha_{0,1,0}$	NA	NA	-3.282	[-4.823;-1.953]	-3.267	[-5.077;-1.814]
$\alpha_{0,1,1}$	NA	NA	2.015	[-1.874;5.589]	1.995	[-1.917;5.917]
$\alpha_{0,1,2}$	NA	NA	NA	NA	0.778	[-4.822;5.626]
$\alpha_{1,0,0}$	NA	NA	-0.266	[-2.303;1.677]	0.326	[-1.943;2.786]
$\alpha_{1,0,1}$	NA	NA	-4.133	[-9.379;-0.826]	-1.320	[-7.438;3.698]
$\alpha_{1,0,2}$	NA	NA	NA	NA	-9.866	[-25.726;-0.192]

Table 2: Posteriori means and 95% credible intervals for model parameters.

NA: not applicable.

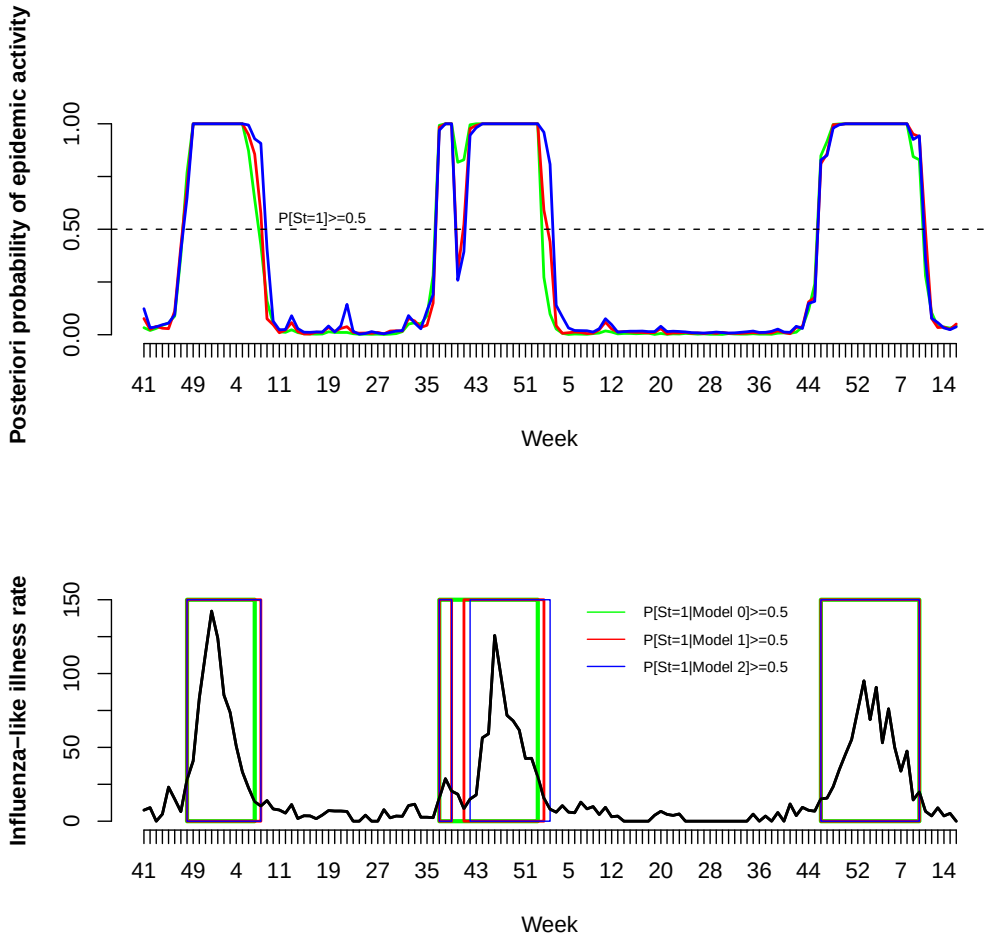

Figure 4: Panel 1: Mean posteriori probabilities of epidemic influenza activity (Model 0: green; Model 1: red; Model 2: blue). Panel 2 : Influenza-like illness rates, reported by Wednesday (solid line); periods of epidemic activity according to model fitted and probability threshold of influenza epidemic activity (colored boxes).

Season	**Model 0**	**Model 1**	**Model 2**
2008-2009	48/08 to 7/09	48/08 to 8/09	47/08 to 8/09
2009-2010	37/09 to 53/10	37/09 to 39/09	37/09 to 39/09
		42/09 to 01/09	42/09 to 02/10
2010-2011	46/10 to 10/11	46/10 to 10/11	46/10 to 10/11

Table 3: Estimated influenza epidemic periods by proposed models for a posterior probability of being in the epidemic state higher than 0.5. Values represent week/year.

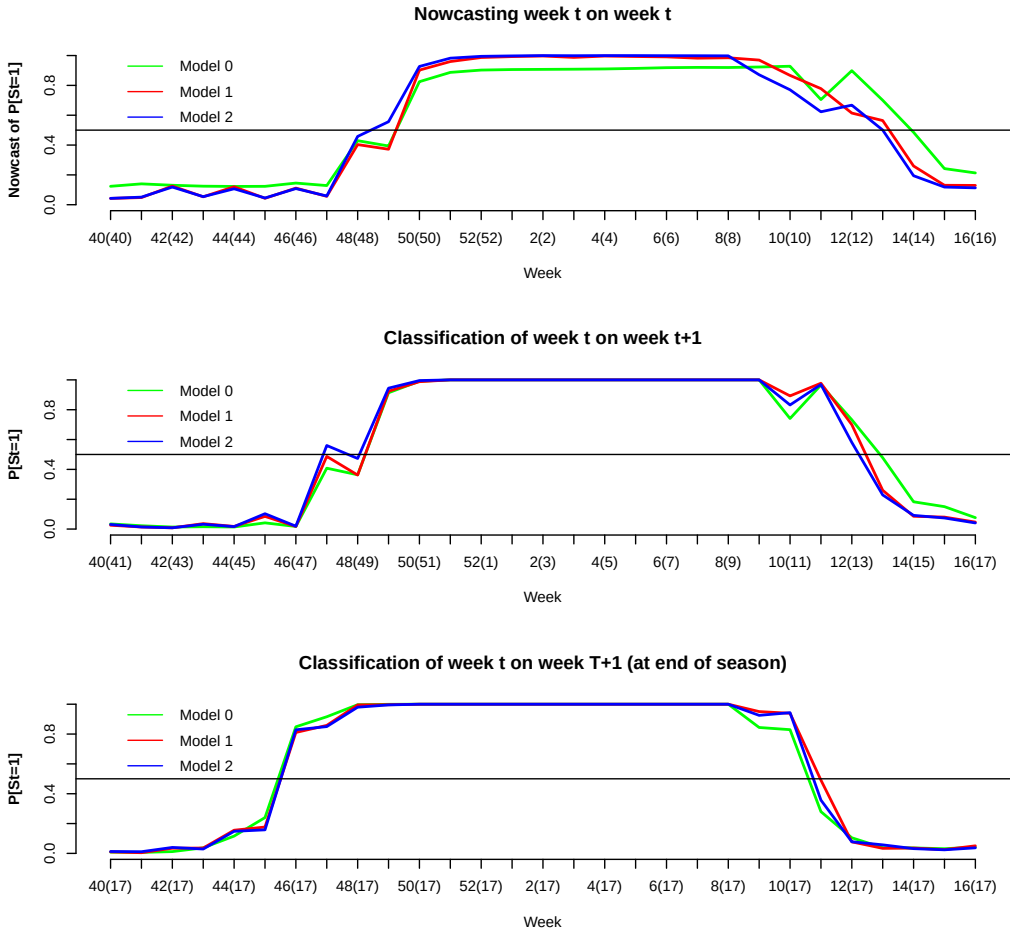

Figure 5: Weekly mean posteriori probabilities of epidemic influenza activity (season 2010-11) calculated Panel 1: in the current week (nowcast); Panel 2: in the following week; Panel 3: at end of the season. (.) week of the calculus.

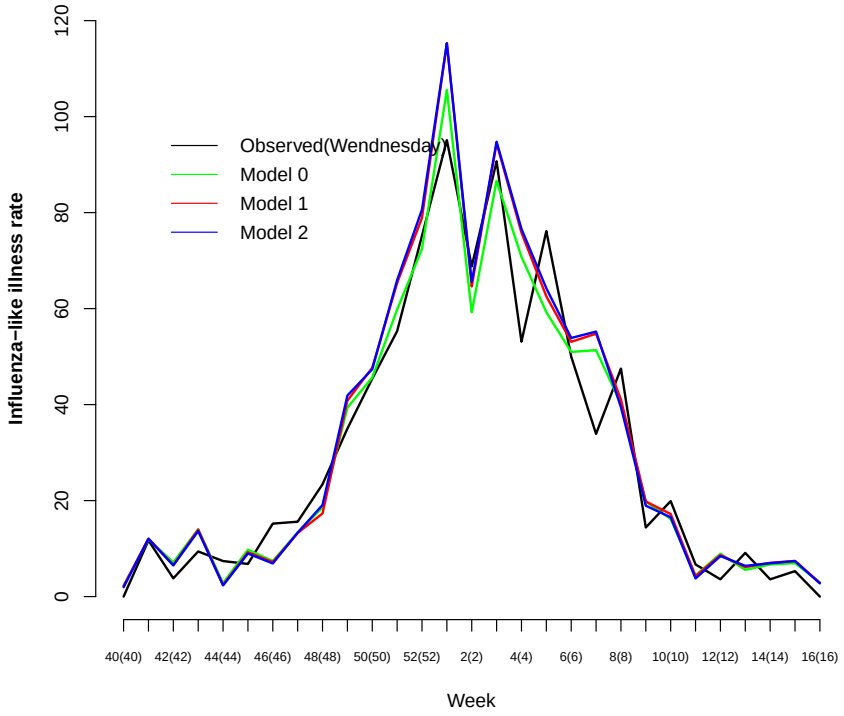

Figure 6: ILI rate nowcast for season 2010-11. (.) week of the calculus.

References

[1] Centre for Disease Control and Prevention. Updated guidelines for evaluating public health surveillance systems: recommendations from the guidelines working group.MMWR 2001; **50** (RR13): 1-35

[2] Centre for Disease Control and Prevention. Framework for evaluating public health surveillance systems for early detection of outbreaks.MMWR 2004; **53** (RR05): 1-11

[3] Lombardo JS, Ross D. Disease Surveillance, a Public Health Priority. In:Lombardo JS, Buckeridge DL, editors. Disease Surveillance. Hoboken, New Jersey: Jonh Wiley and Sons; 2007. p. 1-35.

[4] Fleming DM, van der Velden J and Paget J.The evolution of influenza surveillance in Europe and prospects for the next 10 years. *Vaccine* 2003; **21**: 1749-1753.

[5] European Centre for Prevention and Disease Control. European Influenza Surveillance Network [Internet]. Stockholm: European Centre for Prevention and Disease Control; [cited 2011 May 26]. Available from: http://www.ecdc.europa.eu/en/activities/surveillance/EISN/

[6] Sonesson C, Bock D. A review and discussion of prospective statistical surveillance in public health. Journal of the Royal Statistics Society Series A. 2003; 166: 5-21.

[7] Farrington CP, Andrews N. Outbreak detection: Applications to infectious disease surveillance. In: Brookmeyer R, Stroup DF, editors. Monitoring the Health of Populations: Statistical Principles and Methods for Public Health Surveillance. Oxford: Oxford University Press; 2004. p. 203-231.

[8] Buckridge DL, Burkom HS, Campell M, Hogan WR, Moore A. Algorithms for rapid outbreak detection: a research synthesis. Journal of Biomedical Informatics. 2005; 38: 99-113.

[9] Unkel S, Farrington CP, Garthwaite PH, Robertson C, Andrews N. Statistical methods for the prospective detection of infectious disease outbreaks: a review[Internet]. Milton Keynes, UK: The Open University; 2010 [cited 2011 Jun]. 53 p. Available from:http://stats-www.open.ac.uk/TechnicalReports/OutbreakReviewPaper.pdf

[10] Hohle M. surveillance: An R package for the monitoring of infectious diseases. Computational Statistics. 2007; 22: 37-47.

[11] Deckers JG, Paget WJ, Schellevis FG, Fleming DM. European primary care surveillance networks: their structure and operation. Fam Pract. 2006 Apr;23(2):151-8.

[12] Fleming DM and Elliot AJ. Lessons from 40 years' surveillance of influenza in England and Wales. Epidemiol Infect. 2008 July; 136(7): 866-875.

[13] Truyers C, Lesaffre E, Bartholomeeusen S, Aertgeerts B, Snacken R, Brochier B, Yane F, Buntinx F. Computerized general practice based networks yield comparable performance with sentinel data in monitoring epidemiological time-course of influenza-like illness and acute respiratory illness. BMC Fam Pract. 2010 Mar 22;11:24.

[14] Nicoll A, Ammon A, Amato Gauci A, Ciancio B, Zucs P, Devaux I, Plata F, Mazick A, Mølbak K, Asikainen T, Kramarz P. Experience and lessons from surveillance and studies of the 2009 pandemic in Europe. Public Health. 2010; 124(1):14-23.

[15] Donker T, van Boven M, van Ballegooijen WM, Van't Klooster TM, Wielders CC, Wallinga J. Nowcasting pandemic influenza A/H1N1 2009 hospitalizations in the Netherlands. Eur J Epidemiol. 2011 Mar;26(3):195-201. Epub 2011 Mar 18.

[16] Strat L, Carrat F. Monitoring epidemiologic surveillance data using Hidden Markov Chains models. Statistics in Medicine. 1999; **18** 3463-3478.

[17] Rath TM, Carreras M, Sebastiani P. Automated Detection of Influenza Epidemics. University of Massachusetts. 2003.

[18] Martínez-Beneito MA, Conesa D, Lopéz-Quiléz A, Lopez-Maside A. Bayesian Markov switching models for the early detection of influenza epidemics. Statistics in Medicine. 2008; **27** 4455-4468.

[19] Zucchini W, MacDonald IL. Hidden Markov Models for Time Series: An introduction using R. Boca Raton: Chapman and Hall/CRC; 2009 275p.

[20] Paroli R, Spezia L. Bayesian inference in non-homogeneous Markov mixtures of periodic autoregressions with state-dependent exogenous variables. Computationals Statistics and Data Analysis. 2008; **52**:2311–2330.

[21] Falcão IM, de Andrade HR, Santos AS, Paixão MT, Falcão JM. Programme for the surveillance of influenza in Portugal: results of the period 1990-1996. J Epidemiol Community Health. 1998 Apr;52 Suppl 1:39S-42S.

[22] Chib S. Calculating posterior distributions and modal estimates in Markov mixtures models. Journal of Econometrics. 1996 Apr;75:79-97.

[23] R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[24] Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences (with discussion). Statistical Science. 1992; 7:457-511.

[25] Raftery AE, Lewis S. How Many Iterations in the Gibbs Sampler? In: Bernardo JM, Berger J, Dawid AP, Smith AMF, editors. Bayesian statis-

tics. Oxford: Oxford University Press; 1992. p. 763-773.

[26] Plummer M, Best N, Cowles K, Vines K. CODA: Convergence Diagnosis and Output Analysis for MCMC. R News. 2006; 6(1):7-11.

[27] Chib S. Marginal likelihood from the Gibbs output. Journal of the American Statistical Association. 1995;90 (432): 1313-1321.

[28] Chib S, Jeliazkov I. Marginal Likelihood From the Metropolis-Hastings Output. Journal of the American Statistical Association. 2001;96 (453): 270-281.

[29] Flasche S, Hens N, Boelle PY, Mossong J, van Ballegooijen WM, Nunes B, Rizzo C, Popovici F, Santa-Olalla P, Hruba F, Parmakova K, Baguelin M, van Hoek AJ, Desenclos JC, Bernillon P, Camara AL, Wallinga J, Asikainen T, White PJ, Edmunds WJ. Different transmission patterns in the early stages of the influenza A(H1N1)v pandemic: A comparative analysis of 12 European countries. Epidemics. 2011 Jun;3(2):125-33. Epub 2011 Apr 13.

[30] Departamento de Doenças Infecciosas e Departamento de Epidemiologia. Instituto Nacional de Saude Dr. Ricardo Jorge. Relatório do Programa Nacional da Vigilância da Gripe - epocas 2008-09 e 2009-10 [Influenza Surveillance National Program Report - seasons 2008-09 and 2009-10]. 2010 [Internet]. Lisboa: 2010. Instituto Nacional de Saude Dr. Ricardo Jorge. [cited 2011 Jul]. 61 p. Available

from:http://www.insa.pt/sites/INSA/Portugues/Publicacoes/

Outros/Documents/DoencasInfecciosas/.

Appendix 1

The bayesian inference of the HMM proposed is done by sampling from the posterior distribution (equation 4). The model parameters and the hidden states $\boldsymbol{\Psi} = (\mu, \boldsymbol{\tau}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{s}^T)'$ for the non-homogenous models 1 and 2, and $\boldsymbol{\Psi} = (\mu, \boldsymbol{\tau}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{s}^T)'$ for the homogenous model 0 are numerically estimated using a bayesian approach via Markov chain Monte Carlo (MCMC) methods. For the homogenous model 0 all the parameters are sampled using the Gibbs algorithm, for the non-homogenous models 1 and 2 exception is made for the parameters of the time-variant probabilities transition matrix $\boldsymbol{\alpha}$ that are sampled using a Metropolis-Hastings algorithm. The state sequence $\boldsymbol{s}^T$ is sampled using the *ff-bs* algorithm: forward filtering - backward sampling algorithm.

More specifically, given the vector $\boldsymbol{\Psi}^{(k-1)} = \left(\mu^{(k-1)}, \boldsymbol{\tau}^{(k-1)}, \boldsymbol{\beta}^{(k-1)}, \theta_0^{(k-1)}, \boldsymbol{\theta}_1^{(k-1)}, \boldsymbol{\alpha}^{(k-1)}, \boldsymbol{s}^{T(k-1)}\right)'$ (for the homogenous model $\boldsymbol{\alpha}^{(k-1)}$ is substituted by $\boldsymbol{\gamma}^{(k-1)}$)generated from the parameters and hidden states in the k-1 iteration under the constraint $\tau_0^{(k-1)} > \tau_1^{(k-1)}$, the generic steps to obtain their values in the k iteration are:

1. The state sequence $\boldsymbol{s}^{T(k)}$ is generated by the *ff-bs* algorithm. In this algorithm the filtered probabilities of the states are computed going forward, then the conditional probabilities of the hidden states are computed going backward and used to sample the hidden states from their full conditional distribution suppressing the condition on

$\mu, \boldsymbol{\tau}, \boldsymbol{\beta}, \theta_0, \boldsymbol{\theta}_1, \boldsymbol{\alpha}$ or $\boldsymbol{\gamma}$.

Let us consider the following probabilities vectors:

$$\boldsymbol{\xi}_{t+1/t} = \Big[P(S_{t+1} = 0|y^t_{(+1)}), P(S_{t+1} = 1|y^t_{(+1)})\Big]$$

and,

$$\boldsymbol{\xi}_{t/t} = \Big[P(S_t = 0|y^t_{(+1)}), P(S_t = 1|y^t_{(+1)})\Big]$$

and,

$$\boldsymbol{\xi}_{t} = \Big[P(S_t = 0|S_{t+1}, y^t_{(+1)}), P(S_t = 1|S_{t+1}, y^t_{(+1)})\Big]$$

where $y^t_{(+1)} = (y_{1(2)}, y_{2(3)}, ..., y_{t(t+1)})'$

Given this the *ff-bs* algorithm proceeds with the following steps:

(a)

$$\text{Set} \qquad \boldsymbol{\xi}^{(k)}_{1/0} = \left[\frac{1}{2}, \frac{1}{2}\right]$$

(b) for $t = 1, ..., T-1$

$$\boldsymbol{\xi}^{(k)}_{t/t} = \frac{\boldsymbol{\xi}^{(k)}_{t/t-1}\boldsymbol{F}^{(k-1)}_t}{\mathbf{1}'\left(\boldsymbol{\xi}^{(k)}_{t/t-1}\boldsymbol{F}^{(k-1)}_t\right)'}$$

and,

$$\boldsymbol{\xi}^{(k)}_{t+1/t} = \boldsymbol{\Gamma}^{t(k-1)}\boldsymbol{\xi}^{'(k)}_{t/t}$$

where

$$\boldsymbol{F}^{(k-1)}_t = \begin{bmatrix} f\Big(y_{t(t+1)}|s^{(k-1)}_t = 0\Big) & 0 \\ 0 & f\Big(y_{t(t+1)}|s^{(k-1)}_t = 1\Big) \end{bmatrix}$$

and

$$\Gamma^{t(k-1)} = \begin{bmatrix} \gamma_{0,0}^{t(k-1)} & \gamma_{0,1}^{t(k-1)} \\ \gamma_{1,0}^{t(k-1)} & \gamma_{1,1}^{t(k-1)} \end{bmatrix}$$

(c) For $t = T$ obtain:

$$\boldsymbol{\xi}_{T/T}^{(k)} = \frac{\boldsymbol{\xi}_{T/T-1}^{(k)} \boldsymbol{F}_T^{(k-1)}}{\mathbf{1}' \left(\boldsymbol{\xi}_{T/T-1}^{(k)} \boldsymbol{F}_T^{(k-1)} \right)'}$$

(d) Generate $s_T^{(k)}$ from $\xi_{T/T}^{(k)}$

(e) Obtain from $t = T-1, ..., 1$:

$$\boldsymbol{\xi}_t^{(k)} = \frac{\boldsymbol{\xi}_{t/t}^{(k)} \boldsymbol{\Gamma}'^{t(k-1)}_{\bullet s_{t+1}^{t(k)}}}{\mathbf{1}' \left(\boldsymbol{\xi}_{t/t}^{(k)} \boldsymbol{\Gamma}'^{t(k-1)}_{\bullet s_{t+1}^{t(k)}} \right)'}$$

$\boldsymbol{\Gamma}^{t(k-1)}_{\bullet s_{t+1}^{t(k)}}$ is the column of $\boldsymbol{\Gamma}^{t(k-1)}$ correspondent to the state of $s_{t+1}^{(k)}$. By example if $s_{t+1}^{(k)} = 0$ then

$$\boldsymbol{\Gamma}^{t(k-1)}_{\bullet s_{t+1}^{t(k)}} = \begin{bmatrix} \gamma_{0,0}^{t(k-1)} \\ \gamma_{1,0}^{t(k-1)} \end{bmatrix}$$

(f) Generate $s_t^{(k)}$ from $\boldsymbol{\xi}_t^{(k)}$. At the end of this step the vector of states $\boldsymbol{s}^{T(k)}$ are generated.

2. The precisions $\tau_i, i = 0, 1$ are generated independently from their full conditional distribution which are Gamma distributions with parameters:

$$\alpha_0 = \frac{T_0^{(k)}}{2} + \alpha_\Sigma$$

and

$$\beta_0 = \frac{1}{2} \sum_{\{t \geq 1 : s_t^{(k)} = 0\}} \left(y_{t(t+1)} - \mu^{(k-1)} - \beta_1^{(k-1)} v_{1,t} - \beta_2^{(k-1)} v_{2,t} - \theta_0^{(k-1)} y_{t(t)0} \right) + \beta_\Sigma,$$

to generate $\tau_0^{(k)}$ and

$$\alpha_1 = \frac{T_1}{2}^{(k)} + \alpha_\Sigma$$

and

$$\beta = 1/2 \sum_{\{t \geq 1 : s_t^{(k)} = 1\}} \left(y_{t(t+1)} - \mu^{(k-1)} - \beta_1^{(k-1)} v_{1,t} - \beta_2^{(k-1)} v_{2,t} - \theta_{1,1}^{(k-1)} y_{t(t)1} - \theta_{1,2}^{(k-1)} y_{t(t)1}^2 \right) + \beta_\Sigma,$$

to generate $\tau_1^{(k)}$.

Here the increasing order constrain must be applied if $\tau_0^{(k)} > \tau_1^{(k)}$. In this situation a permutation must applied to the parameter values in order to change the values attributed to the non-epidemic state to the epidemic state and vice versa.

3. The full conditional distribution from which the signal value $\mu^{(k)}$ of the common part is generated from is a Normal distribution with parameters:

$$\mu = \frac{\left[\tau_0^{(k)} \sum_{\{t \geq 1 : s_t^{(k)} = 0\}} \left(y_{t(t+1)} - \beta_1^{(k-1)} v_{1,t} - \beta_2^{(k-1)} v_{2,t} - \theta_0^{(k-1)} y_{t(t)0} \right) \right]}{\left(T_0^{(k)} \tau_0^{(k)} + T_1^{(k)} \tau_1^{(k)} + \tau_M \right)} +$$
$$+ \frac{\left[\tau_1^{(k)} \sum_{\{t \geq 1 : s_t^{(k)} = 1\}} \left(y_{t(t+1)} - \beta_1^{(k-1)} v_{1,t} - \beta_2^{(k-1)} v_{2,t} \theta_{1,1}^{(k-1)} y_{t(t)1} - \theta_{1,2}^{(k-1)} y_{t(t)1}^2 \right) + \mu_M \tau_M \right]}{\left(T_0^{(k)} \tau_0^{(k)} + T_1^{(k)} \tau_1^{(k)} + \tau_M \right)}$$

and,

$$\tau = \left(T_0^{(k)}\tau_0^{(k)} + T_1^{(k)}\tau_1^{(k)} + \tau_M\right)$$

4. The vector of parameters $\boldsymbol{\beta}^{(k)}\left(\beta_1^{(k)}, \beta_2^{(k)}\right)$ are generated from their full conditional distribution which is a Multivariate Normal distribution with parameters:

$$\boldsymbol{\mu} = \left(C'Q^{(k)}V + \Sigma_B^{-1}\right)^{-1}\left(C'Q^{(k)}\tilde{y}^{T(k)} + \Sigma_B^{-1}\mu_B\right)$$

and

$$\Sigma = \left(C'Q^{(k)}C + \Sigma_B^{-1}\right)^{-1}$$

Where $Q^{(k)}$ is the diagonal matrix which the t-th term is $\tau_0^{(k)}$ if $s_t^{(k)} = 0$ and equal to $\tau_1^{(k)}$ if $s_t^{(k)} = 1$. On the other hand $\tilde{y}^{T(k)}$ is the vector with generic terms $\left(y_{t(t+1)} - \mu^{(k)} - \theta_0^{(k-1)}y_{t(t)0}\right)$ if $s_t^{(k)} = 0$ and equal to $\left(y_{t(t+1)} - \mu^{(k)} - \theta_{1,1}^{(k-1)}y_{t(t)1} - \theta_{1,2}^{(k-1)}y_{t(t)1}^2\right)$ if $s_t^{(k)} = 1$.

5. The parameter $\theta_0^{(k)}$ is generated from its full conditional distribution, which is Normal with parameters:

$$\mu = \frac{\tau_0^{(k)}\sum_{\{t\geq 1:s_t=0\}} y_{t(t)0}(y_{t(t+1)} - \mu - \beta_1^{(k)}v_{1,t} - \beta_2^{(k)}v_{2,t}) + \mu_\theta\tau_\theta}{\tau_0^{(k)}\sum_{\{t\geq 1:s_t=0\}} y_{t(t)0}^2 - \tau_\theta}$$

$$\tau = \tau_0^{(k)}\sum_{\{t\geq 1:s_t=0\}} y_{t(t)0}^2 - \tau_\theta$$

On the other hand the parameter $\boldsymbol{\theta}_1^{(k)}$ is generated from a multivariate Normal distribution with parameters:

$$\boldsymbol{\mu} = \left(Y'Q_1^{(k)}Y + \Sigma_\theta^{-1}\right)^{-1}\left(Y'Q_1^{(k)}\tilde{y}^{T(k)} + \Sigma_\theta^{-1}\mu_\theta\right)$$

and

$$\Sigma = \left(Y' Q_1^{(k)} Y + \Sigma_\theta^{-1}\right)^{-1}$$

Where $Q_1^{(k)}$ is the diagonal matrix which the t-*th* term is $\tau_1^{(k)}$ if $s_t^{(k)} = 1$ otherwise 0, $Y = (\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ is $2 \times (T)$ matrix with the covariates $\boldsymbol{Y}_1 = (y_{1(1)1}, ..., y_{t(t)1}, ..., y_{T(T)1})$ and $\boldsymbol{Y}_2 = (y^2_{1(1)1}, ..., y^2_{t(t)1}, ..., y^2_{T(T)1})$.On the other hand $\tilde{y}^{T(k)}$ is the vector with generic terms $\left(y_{t(t+1)} - \mu^{(k)} - \beta_1^{(k)} v_{1,t} - \beta_2^{(k)} v_{2,t}\right)$ if $s_t^{(k)} = 1$ otherwise is 0.

6. Finally for the **homogenous model**: the i-th row of Γ^k matrix is generated from Dirichlet distribution $D(1 + n_{i,0}^{(k)}, 1 + n_{i,1}^{(k)})$, where $n_{i,j}^{(k)}$ is the number of transitions from state i to j in the path of states generated in the k-*th* iteration $s^{T(k)}$ and is obtained by:

$$\sum_{t=2}^{T} = I\left\{s_{t-1}^{(k)} = i, s_t^{(k)} = j\right\}$$

where $I\{.\}$ is an indicator function.

For the **non-homogenous models** the transition probabilities matrix parameters are sampled using a random-walk Metropolis-Hastings algorithm: the $\boldsymbol{\alpha}_{0,1}^{(k)} = \left(\alpha_{0,1,0}^{(k)}, \alpha_{0,1,1}^{(k)}\right)$ and $\boldsymbol{\alpha}_{1,0}^{(k)} = \left(\alpha_{1,0,0}^{(k)}, \alpha_{1,0,1}^{(k)}\right)$ for model 1 and $\boldsymbol{\alpha}_{0,1}^{(k)} = \left(\alpha_{0,1,0}^{(k)}, \alpha_{0,1,1}^{(k)}, \alpha_{0,1,2}^{(k)}\right)$ and $\boldsymbol{\alpha}_{1,0}^{(k)} = \left(\alpha_{1,0,0}^{(k)}, \alpha_{1,0,1}^{(k)}, \alpha_{1,0,2}^{(k)}\right)$ for model 2 are independently generated from a random walk:

$$\boldsymbol{\alpha}_{i,j}^{(k)} = \boldsymbol{\alpha}_{i,j}^{(k-1)} + U_A$$

with $i, j = 0, 1$ and $i \neq j$. Where U_A is generated from Multivariate Normal distribution with parameters $\mu_{\mathbf{A}} = (0, 0)$ and covariance matrix Σ_A. So in each iteration the vectors $\boldsymbol{\alpha}_{0,1}^{(k)}$ and $\boldsymbol{\alpha}_{1,0}^{(k)}$ are accepted respectively with probabilities:

$$A1(\boldsymbol{\alpha}_{0,1}^{(k)}, \boldsymbol{\alpha}_{0,1}^{(k-1)}) = \min\left(1; \frac{\pi\left(\boldsymbol{\alpha}_{0,1}^{(k)} | s^{T(k)}, \boldsymbol{Z}, \delta\right)}{\pi\left(\boldsymbol{\alpha}_{0,1}^{(k-1)} | s^{T(k)}, \boldsymbol{Z}, \delta\right)}\right)$$

and,

$$A2(\boldsymbol{\alpha}_{1,0}^{(k)}, \boldsymbol{\alpha}_{1,0}^{(k-1)}) = \min\left(1; \frac{\pi\left(\boldsymbol{\alpha}_{1,0}^{(k)} | s^{T(k)}, \boldsymbol{Z}, \delta\right)}{\pi\left(\boldsymbol{\alpha}_{1,0}^{(k-1)} | s^{T(k)}, \boldsymbol{Z}, \delta\right)}\right)$$

where $\boldsymbol{Z} = (y_{1(1)}, ..., y_{t(t)}, ..., y_{T(T)})'$ for Model 1 and $\boldsymbol{Z} = (\boldsymbol{y}_0^T, \boldsymbol{v}_{(+1)}^{T-1})$ for Model 2. Here

$$\pi\left(\boldsymbol{\alpha}_{0,1}^{(k)} | s^{T(k)}, \boldsymbol{Z}, \delta\right) \propto$$

$$\propto \delta_{s_1^{(k)}} \prod_{\{t \geq 2 : s_{t-1}^{(k)} = 0\}} \gamma_{s_{t-1}^{(k)}, s_t^{(k)}}^{t(k)} \exp\{-1/2(\boldsymbol{\alpha}_{0,1}^{(k)} - \mu_A)' \Sigma_A^{-1} (\boldsymbol{\alpha}_{0,1}^{(k)} - \mu_A)\}$$

and,

$$\pi\left(\boldsymbol{\alpha}_{1,0}^{(k)} | s^{T(k)}, Z, \delta\right) \propto$$

$$\propto \delta_{s_1^{(k)}} \prod_{\{t \geq 2 : s_{t-1}^{(k)} = 1\}} \gamma_{s_{t-1}^{(k)}, s_t^{(k)}}^{t(k)} \exp\{-1/2(\boldsymbol{\alpha}_{1,0}^{(k)} - \mu_A)' \Sigma_A^{-1} (\boldsymbol{\alpha}_{1,0}^{(k)} - \mu_A)\}$$

where $\gamma_{0,j}^{t(k)}$ and $\gamma_{1,j}^{t(k)}$ with $j = 0, 1$ are functions of $\boldsymbol{Z}$ and also functions of $\boldsymbol{\alpha}_{0,1}^{(k)}$ and $\boldsymbol{\alpha}_{1,0}^{(k)}$, respectively.